Open Access Library Journal
2019, Volume 6, e5935
ISSN Online: 2333-9721
ISSN Print: 2333-9705

An Evaluation of IELTS Speaking Test

Jingyi Li

University of Nottingham, Nottingham, UK
Email: ttxjl90@nottingham.ac.uk

How to cite this paper: Li, J.Y. (2019) An Evaluation of IELTS Speaking Test. *Open Access Library Journal*, **6**: e5935.
https://doi.org/10.4236/oalib.1105935

Received: November 18, 2019
Accepted: December 23, 2019
Published: December 26, 2019

Abstract

In this paper, an evaluation of IELTS Speaking Test was discussed in detail with the support of the brief introduction of purpose of a test, reliability and validity in a test and three eras of language testing. Focusing on examination contexts, both strength and weakness of IELTS Speaking test were discussed from perspectives of validity, reliability and practicality. To conclude, the IELTS speaking test is generally reliable and valid, although there are some controversial elements affecting the degree of reliability and validity, which would be further researched and discussed in this paper. Thereafter, possible suggestions were given following: 1) an intervention; and 2) video-conferencing delivery; and 3) double-marking method. IELTS Speaking Test is one particular module of the IELTS test, which is taken in different locations on the same or different day from listening, reading and writing test in the IELTS. By adopting face-to-face form, the candidate's performance is scored by the examiner only once and the whole process is recorded as a backup. Several weeks later, assessment results will be sent to candidates both online and by mail. In the whole process of testing, subjective elements such as marker performance and candidates' preferences will considerably influence assessment results.

Subject Areas

Education

Keywords

Language Testing, Evaluation, IELTS, Face-to-Face Speaking Test, Video-Conferencing Delivery, Double-Marking Method

1. Introduction

As the high-stakes English test, the International English Language Testing System (IELTS) is well-known as one of the most popular English proficiency tests around the world. Specifically, according to IELTS (http://www.ielts.org/), this testing system measures the language proficiency for the purpose of studying or working in English speaking areas, where English is mainly used for communicating. There are two types of IELTS test provided. Specifically, for people who want to apply for higher education or professional registration, they can apply for IELTS Academic. Otherwise, those wishing to apply for secondary education, training programmes or work experience, may take part in IELTS General Training and enroll in Australia, Canada and the UK.

Assessing English language skills at all levels, the IELTS test has four sections: Speaking, Listening, Reading and Writing. Speaking, as a productive skill, is interactive in nature [1]. Taking interaction into consideration, the IELTS speaking test proceeds in the form of an oral interview in a real-life context. Unlike other sections which are either computer-delivered or paper-based, the speaking section is carried out face-to-face in 11 to 14 minutes with a certificated examiner. Later, results are reported in whole and half bands on a scale from 1 to 9. The IELTS speaking test requires candidates to have integrated speaking ability assessed by four criteria: fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation.

Although commonly accepted by institutions, enterprises and governments around the world, this kind of widespread use makes the IELTS speaking test controversial in the meantime. In other words, whether test results can reflect test takers' language behaviours correctly during the test and whether these behaviours can reflect test takers' real language competence are essential to not only test takers themselves but also employing units who measure employees' competence by test results directly. Thus, it is necessary to find out whether IELTS speaking test is reliable.

This essay comprises six parts. After this brief introduction, this paper will briefly illustrate the theoretical information concerning language testing and introduce the basic information of IELTS speaking test in the literature review section. Thereafter, section three offers an outline of the context, as regards candidates, examiners, test conditions, test structure, rating scale and the development of the IELTS speaking test. Then, section four evaluates the validity, reliability and practicality of the IELTS speaking test supported by the Literature in Section two. Section five contains suggestions made about three aspects: the effect of an intervention, video-conferencing delivery and double-marking methods. Finally, the conclusion provides an overview of key findings of this paper.

2. Literature Review

2.1. The Purpose of Test

As one form of assessment, language testing fulfills various and diverse functions in both the classroom and society [2]. Admittedly, different tests with different functions will have different purposes [3]. Our considered judgements towards language tests should take both the historical evolution of testing and assessment and the legitimate roles of testing in egalitarian societies into consideration [2]. Furthermore, finding out the purpose of testing is fundamental in the process of evaluating practical applications [4]. Here is a basic introduction of four main tests concluded by Hughes [3]: proficiency test, achievement test, diagnostic test and placement test.

Regardless of previous learning and training experience, proficiency tests are designed to find out whether candidates have sufficient command of the language to be considered proficient for a particular purpose [3]. Most proficiency tests are set to show whether candidates have reached a certain standard with respect to a set of specified abilities [5]. Furthermore, standardised proficiency tests tend to be criterion-referenced though not the opposite that is, norm-referenced [6]. That is to say, in the proficiency test, a candidate's performance is tested by the rating criterion standard rather than comparing against other candidates taking the same test [2]. Moreover, demonstrating the four main language skills (namely listening, speaking, reading and writing), proficiency tests are provided for people of all levels, as long as candidates are willing to take these tests [7]. IELTS may, therefore, be considered a proficiency test.

In contrast, achievement tests are directly related to how individual students and groups achieve language courses [3]. Similarly, regardless of language abilities candidates gained before, achievement tests mainly test learners' knowledge at the time of the test [2]. There are two kinds of achievement tests: final achievement tests and progress achievement tests. Learners take final achievement tests at the end of the courses and progress achievement tests in the middle of the course [3].

As for the diagnostic test, this is commonly used for diagnosing students' areas of strength and weakness so as to choose appropriate types and levels of teaching and learning activities [8]. Importantly, diagnostic tests predict primarily what learning still needs to take place [3]. Furthermore, good diagnostic tests are extremely useful for self-instruction [9]. Existing gaps in the command of language may be shown in the results of diagnostics tests, however, a tremendous amount of work and test developers' willingness are needed to produce a diagnostic test, resulting in the hard implementation of a diagnostic test [3].

Similar to the diagnostic test, a placement test may be regarded as a broadband diagnostic test since it distinguishes students from relatively weak to relatively strong in order to form appropriate groups [2]. As the name suggests, placements tests will help educators place students at the stage of teaching programme related to students' abilities [7]. In designing a placement test, the theory of language proficiency or the objectives of the syllabus may be taken into consideration [3].

2.2. Reliability

Reliability is an absolutely essential quality of tests as well as a function of the consistency of scores from one set of tests and test tasks to another [7]. Further, Chalhoub-Deville and Turner [4] describe reliability as "the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure". Generally, reliable tests can evaluate individual's unchanged ability consistently no matter the time or place the test is taken [3]. The more extensive the testing procedure and maker variability is realized, the more reliable a test is [5].

Other than language abilities that language tests want to measure, factors which are largely unsystematic and hence unpredictable, such as poor health, lack of motivation, and test-wiseness affect the test performance and the reliability of language scores [10]. Also, Bachman, Palmer and Palmer [7] indicate that the differences in testing conditions, fatigue and anxiety may affect candidates' performance, which may lead to inconsistent scores from one occasion to another. Hence, identifying the different sources of measurement error should be put into the primary stage [11]. Realizing the effect of these factors, researchers can accordingly minimize measurement error and maximize reliability [8]. Similarly, the following issues should be taken into consideration to increase reliability of tests: the conditions under which the test is taken, psychometric properties embodied in the difficulty indices of test tasks, and standard error of measurement especially near passing scores [4].

With the exception of the first measurable quality of test usefulness—reliability [5], another interrelated concept—validity is equally worth interpreting in relation to language testing (see next section). Indeed, reliability is a necessary condition for validity only because unreliable test scores hardly promise valid interpretation and use [7]. Moreover, reliability relates to the minimum effects of measurement error while validity relates to the maximum effects of the language ability measured [8]. Then, the definition of validity and related issues will be presented as follow.

2.3. Validity

According to Hughes [3], if a test measures accurately what it intends to measure, then it is valid. Further, Messick [12] describes validity as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores". Then, a series of questions concerning which kind of scores are useful and how these scores may help make decisions and whether tests have positive consequences for test takers. Come along the raise of the definition of validity [2]. To make test scores meaningful, demonstration concerned with affective factors should not include the testing ability [8]. There are several subordinate forms of validity which are increasingly used in the evaluation of language tests: content validity and criterion-related validity, face validity.

Firstly, content validity requires a test with a representative sample of the language skills and structures meant to be concerned or covered [7]. The basis of figuring out whether a test contains content validity is a comparison of test specification and test content [11]. On the one hand, content validity is indispensable due to its guarantee of accurate measurement; on the other hand, a harmful washback effect may be caused if the absence of some areas in language tests leads to the ignorance of those knowledgeable in teaching and learning [3].

Secondly, in contrast with content validity, criterion-related validity is established on the basis of empirical correlation between the test scores and criterion scores [12]. There are two kinds of criterion-related validity: concurrent validity and predictive validity. Specifically, concurrent validity is used to illustrate the consistent relationship between scores from a new measurement and a well-established one [8]. Nevertheless, predictive validity refers to the forecasting ability of a test that can predict candidates' future performance [3].

Lastly, if a test looks as if it measures as expected, then this test is said to have face validity [3]. Importantly, face validity reflecting test appearance and test appeal, plays an indispensable role in the acceptability of tests to both test takers and test users considerably [2]. However, face validity is not a scientific notion and cannot provide evidence for construct validity. Necessary as face validity is, teachers and education authorities should consider seriously about plausible tests [8].

2.4. The Different Eras of Language Tests

Spolsky [13] identified three main periods in the development of language testing: pre-scientific period, psychometric-structuralist period and psycholinguistic-sociolinguistic period. Importantly, each language testing period was connected closely to corresponding development of pedagogy, psychology and sociology at that period. Furthermore, these three trends co-exist in time and approach.

More precisely, pre-scientific period was influenced by philology and the study of Latin. Learners learned English language mainly through parsing, translating and remembering prescriptive grammar rules [11]. Hence, assessing the learners' ability of grammar and lexical resources became the main purpose to judge the proficiency of learners in the pre-scientific period. There was no notion of reliability and validity at all at this stage.

However, in the 20th century, linguistics and psychology developed with a flourish. At first, scholars in linguistics who were in favor of structuralism and those in psychology who were in favor of behaviorism viewed language learning as repetitive structure drills of sentence patterns derived from "native speaker" "standard" use. Later, another advocation prevailed, which described language as inborn rules. Since then, correct input became the main focus.

Then, at the psycholinguistic-sociolinguistic stage, Canale and Swain [6] proposed communicative competence. Language testing in this stage was influenced by sociolinguistics, functional linguistics and socio-cultural psychology. Importantly, learners were emphasized as social beings and language was a tool of social exchange. The focus of this era is on the integrative testing containing principles of communicative competence. Hence, the purpose of the test is transformed to examine the target language use with different functions in real life [3].

The IELTS speaking test, as one of the contemporary proficiency tests, has similar features as the tests in post-scientific era. The IELTS speaking test measures candidates' oral language proficiency taking rating criteria as standard. Spoken language production is based on social interaction containing communicative language skills. In addition, candidates have adequate opportunities to speak at length and display their ability in English language meaningfully [14]. Essentially, the IELTS speaking test is considered valid and reliable in the principles of the post-scientific era.

3. Context

According to IELTS Syllabus 2019, the test is supported and undertakenby British Council and IDP: IELTS Australia. The speaking test, as one of four sections, seeks to assess a wide range of skills in how well the test-taker can: communicate opinions and information on everyday topics naturally; speak at length according to a given topic appropriately; organise ideas coherently; express and justify opinions; analyse and discuss issues fluently (http://www.ielts.org/). Here is basic information concerning IELTS speaking test.

3.1. Candidates

Scores of IELTS are largely accepted by more than 10,000 organizations globally (http://www.ielts.org/). As long as people want to live, work or study in countries such as Australia, Canada, New Zealand and the United Kingdom, they have to submit eligible IELTS grades. The official website provides test detailed information. It is very easy to navigate around and useful for users registering and paying for the IELTS. The examination expense (around GBP170) should be payed after submitting the information sheet concerning personal information, test types and test timetable. Later, emails and messages will be sent to participants telling them when they will take the test and how to complete the whole procedure. Even some teaching and learning videos to prepare well for IELTS are attached for free to the official website.

According to demographic statistics 2017, almost 78 percent of candidates attend academic IELTS. In the speaking part, participants in Greece, Germany and Canada achieve well at band 7. Average score of speaking tests for both female and male is over band 6.

3.2. Examiners

There are 7000 plus IELTS examiners whose marking performance is monitored and maintained through the IELTS Professional Support Network, designed and managed by British Council and IDP: IELTS Australia. Proficient as these examiners are, regular training sessions and recertification evaluation are required for each of them.

In the speaking section, examiners are dispatched randomly and confidentially. Before the speaking test commences, one examiner will wait for the corresponding candidate dispatched already in the testing place. Importantly, the candidate's performance is scored by the examiner only once and the whole process is recorded as a backup. To maximize test reliability and validity, examiners must follow a script and stick to the rubrics.

3.3. Test Conditions

The speaking test is taken in different locations on the same or different day from listening, reading and writing test in the IELTS. A spacious and quiet room will be provided separately for each test taker and the corresponding examiner. If candidates have any problems, they can turn to invigilators for help in any occasion.

3.4. Test Structures

The content and structure of speaking module of Academic IELTS tests and General IELTS tests are the same. There are three main parts in the IELTS speaking test. Each part fulfils a specific function in terms of interaction pattern, task input and candidate output. A sample of examination process is attached in **Appendix 2**.

Part 1 (introduction): the examiner greets and verifies identification first. Then examiner will ask general daily questions on familiar things such as homes/families, jobs/studies, interests, life routine, friends and so on. This part lasts 4 to 5 minutes.

Part 2 (individual long turn): a task card will be given to talk about a specific topic. Candidates can prepare for 1 minute before speaking at length for no more than two minutes. And taking notes on white board provided by the examiner is permitted. The examiner will not be interrupted during speaking only if the time is up. This part lasts 3 to 4 minutes.

Part 3 (two-way discuss): the examiner will ask two or three or more topic-related questions based on the previous interpretation.

More abstract issues and ideas can be discussed in this part between the examiner and candidate. This part lasts 4 to 5 minutes.

3.5. Rating Scale

The score range of the IELTS speaking tests is from 1 to 9, just like other sections of the IELTS tests. Four sections will be taken into consideration when examiners score the performance of candidates: fluency and coherence, lexical resource, grammatical and accuracy and pronunciation. However, no minimum score is requested to pass the exam. Detailed information is attached in the appendix (speaking: band descriptors—public version).

3.6. The Development of the IELTS Speaking Tests

The IELTS speaking tests, as one module of the IELTS test, were established in 1989 and administered by the British Council, Cambridge English Language Assessment and the International Development Program of Australian universities and colleges. Later, IELTS was revised significantly in 1995. To alleviate the burden of scheduling the speaking assessment, candidates were allowed to take the speaking test on a different day from the other three modules. In 2001, some detailed changes presented in both test content and evaluation part. More specifically, tasks existed in the speaking test and examiner scripts were combined with scoring criteria into the speaking paper [14].

3.7. Research Process

In order to get a good command of general information of IELTS Speaking Test, the researcher visited the official IELTS website and collected data of content, construction, scoring criteria and preparation recommendations of IELTS Speaking Test. Moreover, the researcher read a great amount of series of IELTS Research Reports online. By taking notes and analyzing various research results of other studies, the researcher got conclusions in order and reintegrated points of view based on the solid foundation of literatures to deal with the main research question: is the IELTS Speaking Test reliable?

4. Evaluation

4.1. Validity

4.1.1. Content Validity

As mentioned in the literature review, content validity is concerned with the relevance of the test content to the content of a specific behavioral domain of interest and about the representativeness that item or task content covers [3]. Importantly, identifying the domain specification provides the means for examining relationships between the test performance and performance in other contexts [3]. However, the content validity itself is not sufficient evidence for validity due to the ignorance of how test takers perform [8].

IELTS speaking test with various tasks has content validity, matching the communicative requirements of the test. Speech functions like comparing, summarizing, explaining, suggesting, contrasting, narrating, paraphrasing and analyzing occur regularly in a candidate's output without the influence of external test structure [15].

However, a test may not be completely valid. Questioning skill is absent in the IELTS speaking test [16]. A series of topic-based question-answer adjacency pairs hardly provides candidate with a topic shift or introduction. That is to say, candidates have little opportunity to display their ability to manage topic development and turn-taking [16]. Further, regardless of whether candidates of all levels choose to take the speaking test for general or academic purpose, the content and structure of the speaking test are the same. Admittedly, adolescents who complete the IELTS test for higher educational purposes, speak from experience of a different context from older candidates, who may be here for immigration purposes. Hence, inappropriate topics such as a business-related task may not suit adolescents [17].

4.1.2. Predictive Validity

As shown in the literature review, predictive validity refers to the extent to which test scores predict future behavior [2]. It scores on the predicted behaviour on a criterion which is expected to happen in the future. The relationship between the ability the test appears to measure and the performance predicted plays an indispensable role in the assessment of predictive ability [15]. However, according to Quaid [17], language production in the IELTS tasks (micro level) may not necessarily indicate the overall language adequacy (macro level). Due to the same structures and content in general and academic oral tests, assessing general speaking ability in general context is easier than that in an academic context [15].

Besides, in the whole IELTS testing process, examiners pursue strictly the interactive and close to real-life purpose as the L2 classroom discourse requires [17]. However, this kind of institutional discourse sometimes fails to predict candidates' future professional success. The speculative conclusion which may be gained from the results of speaking tests hardly contribute to the predictive validity [16].

4.1.3. Face Validity

According to Hughes [3], face validity is concerned with the surface credibility or public acceptability of a test. In other words, if a test looks like what it is supposed to measure, this test has face validity. Admittedly, IELTS speaking test meets the criterion of face validity. As mentioned in the literature review, there are two factors influencing face validity of a test: the familiarity of test format and authenticity in test task [15]. Therefore, the analysis of face validity of IELTS speaking test will follow this section as discussed below.

Firstly, the format of IELTS speaking test is quite clear and well established [14]. According to the context, there are three parts: introduction, individual long turn and two-way discussion. The testing flow, testing procedures and testing content and structure in separate parts are all clear for examination takers. Therefore, the IELTS speaking test overall "looks" reliable.

Secondly, examination prescription and related educational resources are easily available for all people including candidates, educators and even parents [18]. Specifically, information concerning what is IELTS, test construction, suitable guidelines and teaching and research is provided on the official website (http://www.elts.org/). Furthermore, the information on the website is not fixed but keeping pace with the times. Related research aimed at assessing IELTS test for better development is collected in the IELTS Research Note (http://www.cambridgeenglish.org/).

4.2. Reliability

IELTS, a non-certificated testing system, do not provide candidates with a pass or fail mark. According to the IELTS (2017), experimental and generalizable studies conclude that although scoring reliability or statistical significance is unavailable, IELTS speaking test has a relatively high correlation coefficient. Two factors affecting the reliability of a test will be analyzed as follows.

4.2.1. Marker Variability

Marker variability in IELTS speaking test which is a subjectively scored test is the inevitable outcome of rating procedure. From a rater perspective, inter-rater and intra-rater are two types of raters' grading [15]. Based on theory in the aforementioned literature review, inter-rater reliability refers to the consistency of different raters agreeing on the same performances, while intra-rater reliability refers to the consistency of the same rater repeating the same performance by applying the same criteria. Inevitably, IELTS speaking test is influenced largely by inter-rater reliability due to the single rater.

Given this situation of a single rater rating, interviewer's variability and subjectivity decide a candidate's reported proficiency level which may be not accordance with his/her inherent ability [15]. Brown and Hill [19] distinguish test raters into two types: "the difficult interviewers and easy interviewers". Specifically, difficult interviewers pay great attention to complex skills of speculating and justifying opinions. For example, they may interrupt candidates with another critical question before they complete the previous task. On the other hand, easy interviewers tend to question in an easy and economical way and seldom bother test takers with argumentative questions. In addition, they may choose open questions and present scaffolding behaviour. Furthermore, cultural expectations under test conditions play an important role in the reliability of the test. Examiners in different cultures may focus on different aspects of language production [2].

4.2.2. Test Conditions

The test conditions may affect the test results and test reliability. Physical environment, partner compatibility and test procedure are main factors contributing to the reliability of a test [15]. As a whole, the effect of the physical environment in IELTS speaking test may be both beneficial and harmful. The indoor testing environment is not affected by external weather factors that may affect the reliability of the test. In addition, separate and closed rooms provide test takers with a quiet testing atmosphere, isolating noisy voices of the outside world [7].

However, face-to-face test conditions in real and limited time may lead to examiners' tension and anxiety between an interlocutor and a candidate especially when candidates cannot hear from the examiner at the first time or feel it hard to complete challenging tasks [20]. Besides, eye contact and close social distance are inevitable in face-to-face test conditions. Scores of candidates who prefer to stand relatively far apart and be too shy to look directly, may be influenced by examiners who may view these behaviors as dishonest or disrespectful.

Considering the aspect of test procedure, interview format may be the only way to assess the speaking skill in the IELTS speaking test. Not every person can perform well and feel comfortable in this formal and somewhat restricted context. For instance, the same testing taker may perform well in role-play tasks but behave terribly in the fixed answering question mode. As Hughes [3] suggests that "the addition of further items will make a test more reliable".

4.3. Practicality

Bachman and Palmer [10] viewed "usefulness" as a superordinate conclusion containing reliability, authenticity, interactiveness and practicality. Practicality, as one of these qualities, is indispensable in the evaluation of a test. Importantly, practicality is concerned with economy, administration, scoring and interpretation of results [15]. Furthermore, practicality is related to the implementation of the test rather than the meaning of test scores [6]. In addition, of equal importantance is that a balanced cyclical model is the guarantee for a useful test, in which reliability, authenticity, interactiveness and practicality effect equally [5].

Admittedly, IELTS speaking test is comparatively practical because of its ease of administration. IELTS test has been put into practice for almost 30 years so that examination process and precautions, examiner training system and evaluation researches are deployed widely and systematically. Moreover, the short testing time and brief procedures improve the practicability of IELTS speaking test. Participants taking the speaking test for just 11 to 14 minutes will maintain their freshness and reduce the fatigue factor.

However, there is no absolutely practical test. Admittedly, the continued use of IELTS oral proficiency interviews wastes unnecessary human, material and time resources. Compared to IELTS oral proficiency interviews, candidates have a highly positive attitude towards a computer-based speaking test mode of delivery being less threatening and anxious [5].

To conclude, IELTS speaking test has relatively high practicality due to its short time-consuming and easy administration, although it still has its deficiencies inevitably.

5. Suggestions

5.1. The Effect of an Intervention

Based on evaluation in Section 4, IELTS speaking test cannot be completely valid. Although speech functions like comparing, summarizing, explaining, suggesting, contrasting, narrating, paraphrasing and analyzing emerge regularly in a candidate's output during the test, candidates hardly use their questioning skills in any of the three testing parts [16].

A suggestion in this situation is to add an intervention lasting two minutes, namely the addition of a fourth part after the two-way discuss. Candidates questioning examiners could start in various ways. For example, the candidate might ask the examiner questions concerning previous topics discussed in Sections 2 and 3. Further, test takers can ask a follow-up question based on the leading statement that examiners introduce firstly.

The significance of the added section is to provide a more active role for candidates who apply for IELTS mainly for higher education purposes due to the high similarity of questioning context between IELTS speaking test and group settings in universities [20]. Moreover, candidates questioning pattern may provide a more naturalistic, two-way interaction which may originally occur in the previous section. Also, raters can confirm decisions on grades comprehensively by taking data from candidates questioning [16].

5.2. Video-Conferencing Delivery

In the previous evaluation section, test conditions such as eye contact and close range may influence test takers' performance so as to affect final results. Emotional factors including tension and anxiety caused by physical testing conditions may lead to unexpected results especially for those who are not good at dealing with emergent issues [7]. Finally, due to the fixed testing place, test takers in remote districts need to devote energy, time and finance to take a proficiency test.

A suggestion to this is to carry out the video-conferencing delivery mode. According to the study of Nakatsuhara *et al.*, [20], functional output, examiner interviewing, and rating behaviors change in the video-conferencing delivery mode compared to standard face-to-face behaviour. For example, the increased use of negotiation and signal shows test takers' engagement and understanding in the communication under the video-conferencing mode.

The significance of this is that, by using the video-conferencing delivery, test takers may communicate with raters in real time through audio and video in two or more locations, which offers the practical advantage of connecting candidates and examiners who are continents apart (ibid.).

5.3. Double-Marking Methods

As shown in the evaluation section, marker variability which is inevitable in the face-to-face test structure is an indispensable element in assessing the reliability of a test [15]. Admittedly, IELTS speaking test is influenced largely by inter-rater reliability due to the single rater. Hence, factors like different cultural expectations and subjective preferences of a single rater lead to a low reliability of the test.

A suggestion in this situation is to arrange an examiner to do a "live" rating during the test sessions and another examiner double-mark the audio or video recorded test session later. Compared to the large-scale test operationalization and high costs in double marking several times in the real-life context, this kind of double-mark-record system seems more practical. Moreover, rapid advances in computer technology in the present age make gathering and transmission of recorded performances easily available (ibid.)

The significance of non-live double-marking methods is that using video or audio recordings of the candidates' spoken performance may help examiners notice both negative and positive features that may have been missed in the one-off interview and several markers based on different raters strengthen the reliability of a test (ibid.). Therefore, double-marking methods make scores to some extent more reliable.

6. Conclusion

6.1. Overview

The purpose of this study has been to assess the reliability and validity of IELTS speaking test. The study has presented a critical review of key literature on the purpose of tests, the basic information about reliability and validity, the different eras of language assessment and the key features of the post-scientific era. Moreover, this paper evaluates the content validity, predictive validity and face validity and marker reliability and test conditions and practicality based on the context of IELTS speaking test and the literature. Then, suggestions are given from three aspects: the effect of an intervention, video-conferencing delivery and double-marking methods.

6.2. Summary of Key Findings

The IELTS speaking test is generally reliable and valid. Specifically, the test has validity in terms of content validity and face validity. The test is reliable on the content due to the brief complementation of three sections communicatively without related academic domains. The test is reliable on the appearance since the accessibility of the format and related information about the test. Candidates can take the IELTS speaking test under adequate preparation based on educational materials and resources.

However, a test cannot be totally valid. In terms of content validity, questioning skill is absent contributing to hard production of a topic shift or introduction. Further, the IELTS speaking test has the same content and structure for candidates from all levels, although not suitable in any occasion. Moreover, the institutional discourse sometimes fails to predict candidates' future professional success because of the speculative conclusion from examiners. Finally, on the part of test conditions and marker variability, the IELTS speaking test needs to be improved as well.

Finally, three suggestions are provided as possible thoughts based on the evaluation content. Firstly, a two-minute-intervention can be added in the fourth part following the two-way discuss section. Secondly, video-conferencing delivery mode can be taken into consideration. Thirdly, double-marking methods whether in audio or video are practicable to increase the reliability and validity.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Hughes, R. (2011) Teaching and Researching Speaking. 2nd Edition, Routledge, London. https://doi.org/10.4324/9781315833736

[2] Fulcher, G. (2010) Practical Language Testing. Glenn Fulcher, London.

[3] Hughes, A. (2003) Testing for Language Teachers. 2nd Edition, Arthur Hughes, Cambridge. https://doi.org/10.1017/CBO9780511732980

[4] Chalhoub-Deville, M. and Turner, C.E. (2000) What to Look for in ESL Admission Tests: Cambridge Certificate Exams, IELTS, and TOEFL. *System*, **28**, 523-539. https://doi.org/10.1016/S0346-251X(00)00036-1

[5] Quaid, E. (2018) Reviewing the IELTS Speaking Test in East Asia: Theoretical and Practice-Based Insights. *Language Testing in Asia*, **8**, 1-9. https://doi.org/10.1186/s40468-018-0056-5

[6] Fulcher, G. and Davidson, F. (2007) Language Testing and Assessment. Glenn Fulcher and Fred Davidson, London.

[7] Bachman, P. and Palmer, A.S. (2010) Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World. Lyle Bachman and Adrian Palmer, Oxford.

[8] Bachman, L.F. (1990) Fundamental Considerations in Language Testing. Lyle F. Bachman, Oxford.

[9] Alderson, J.C., Clapham, C. and Wall, D. (1995) Language Test Construction and Evaluation. Caroline Clapham and Dianne Wall, Cambridge.

[10] Bachman, P. and Palmer, A.S. (1996) Language Testing in Practice: Designing and Developing Useful Language Tests. Lyle F. Bachman and Adrian S. Palmer, Oxford.

[11] Davies, A. (1978) Language Testing. *Language Teaching*, **11**, 145-159. https://doi.org/10.1017/S0261444800003748

[12] Messick, S. (1989) Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher*, **18**, 5-11. https://doi.org/10.3102/0013189X018002005

[13] Spolsky, B. (1977) Language Testing: Art or Science in Nickel, G (ED). *Proceedings of the Fourth International Congress of Applied Linguistics*, **3**.

[14] Taylor, L. (2001) Revising the IELTS Speaking Test: Developments in Test Format and Task Design.

[15] Karim, S. and Haq, N. (2014) An Assessment of IELTS Speaking Test. *International Journal of Evaluation and Research in Education*, **3**, 152-157.

[16] Seedhouse, P. and Morales, S. (2017) Candidates Questioning Examiners in the IELTS Speaking Test: An Intervention Study. IELTS Research Reports Online Series, 43.

[17] Seedhouse, P. and Harris, A. (2011) Topic Development in the IELTS Speaking Test. *IELTS Research Reports*, **12**, 1.

[18] Fernandez, C.J. (2018) Behind a Spoken Performance: Test Takers' Strategic Reactions in a Simulated Part 3 of the IELTS Speaking Test. *Language Testing in Asia*, **8**, 18. https://doi.org/10.1186/s40468-018-0073-4

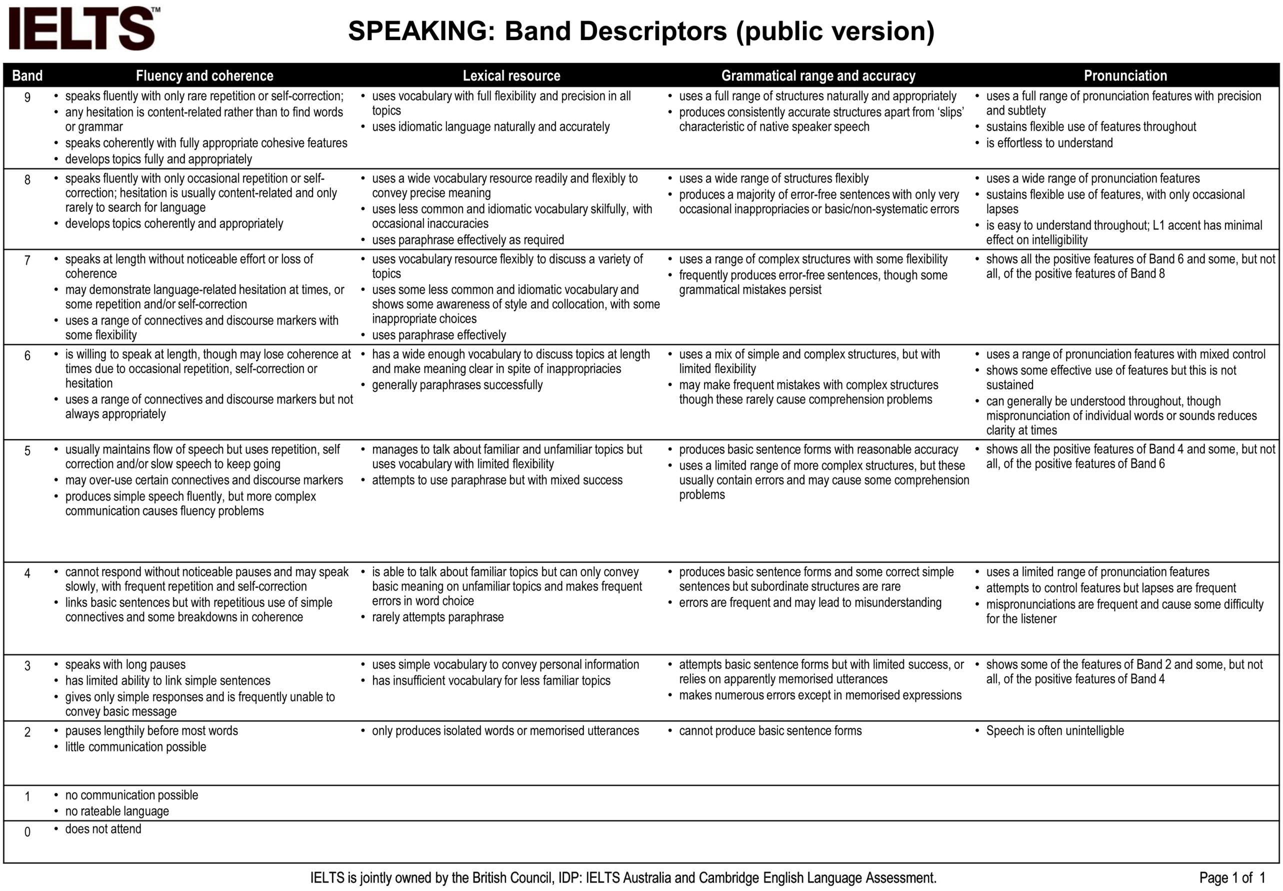
[19] Brown, A. and Hill, K. (1998) Interviewer Style and Candidate Performance in the

IELTS Oral Interview. *International English Language Testing System* (*IELTS*) *Research Reports*, **1**, 1.

[20] Nakatsuhara, F., Inoue, C., Berry, V. and Galaczi, E.D. (2016) Exploring Performance across Two Delivery Modes for the Same L2 Speaking Test: Face-to-Face and Video-Conferencing Delivery: A Preliminary Comparison of Test-Taker and Examiner Behaviour. The IELTS Partners, British Council, Cambridge English Language Assessment and IDP, IELTS Australia.

Appendices

Appendix 1



IELTS™

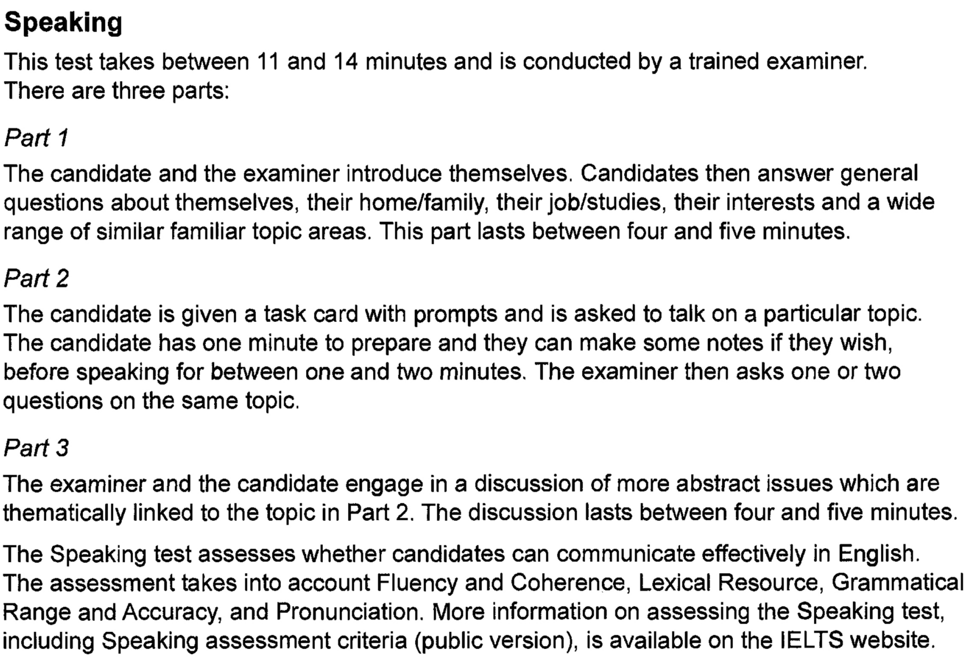
SPEAKING: Band Descriptors (public version)

Band	Fluency and coherence	Lexical resource	Grammatical range and accuracy	Pronunciation
9	• speaks fluently with only rare repetition or self-correction; • any hesitation is content-related rather than to find words or grammar • speaks coherently with fully appropriate cohesive features • develops topics fully and appropriately	• uses vocabulary with full flexibility and precision in all topics • uses idiomatic language naturally and accurately	• uses a full range of structures naturally and appropriately • produces consistently accurate structures apart from 'slips' characteristic of native speaker speech	• uses a full range of pronunciation features with precision and subtlety • sustains flexible use of features throughout • is effortless to understand
8	• speaks fluently with only occasional repetition or self-correction; hesitation is usually content-related and only rarely to search for language • develops topics coherently and appropriately	• uses a wide vocabulary resource readily and flexibly to convey precise meaning • uses less common and idiomatic vocabulary skilfully, with occasional inaccuracies • uses paraphrase effectively as required	• uses a wide range of structures flexibly • produces a majority of error-free sentences with only very occasional inappropriacies or basic/non-systematic errors	• uses a wide range of pronunciation features • sustains flexible use of features, with only occasional lapses • is easy to understand throughout; L1 accent has minimal effect on intelligibility
7	• speaks at length without noticeable effort or loss of coherence • may demonstrate language-related hesitation at times, or some repetition and/or self-correction • uses a range of connectives and discourse markers with some flexibility	• uses vocabulary resource flexibly to discuss a variety of topics • uses some less common and idiomatic vocabulary and shows some awareness of style and collocation, with some inappropriate choices • uses paraphrase effectively	• uses a range of complex structures with some flexibility • frequently produces error-free sentences, though some grammatical mistakes persist	• shows all the positive features of Band 6 and some, but not all, of the positive features of Band 8
6	• is willing to speak at length, though may lose coherence at times due to occasional repetition, self-correction or hesitation • uses a range of connectives and discourse markers but not always appropriately	• has a wide enough vocabulary to discuss topics at length and make meaning clear in spite of inappropriacies • generally paraphrases successfully	• uses a mix of simple and complex structures, but with limited flexibility • may make frequent mistakes with complex structures though these rarely cause comprehension problems	• uses a range of pronunciation features with mixed control • shows some effective use of features but this is not sustained • can generally be understood throughout, though mispronunciation of individual words or sounds reduces clarity at times
5	• usually maintains flow of speech but uses repetition, self correction and/or slow speech to keep going • may over-use certain connectives and discourse markers • produces simple speech fluently, but more complex communication causes fluency problems	• manages to talk about familiar and unfamiliar topics but uses vocabulary with limited flexibility • attempts to use paraphrase but with mixed success	• produces basic sentence forms with reasonable accuracy • uses a limited range of more complex structures, but these usually contain errors and may cause some comprehension problems	• shows all the positive features of Band 4 and some, but not all, of the positive features of Band 6
4	• cannot respond without noticeable pauses and may speak slowly, with frequent repetition and self-correction • links basic sentences but with repetitious use of simple connectives and some breakdowns in coherence	• is able to talk about familiar topics but can only convey basic meaning on unfamiliar topics and makes frequent errors in word choice • rarely attempts paraphrase	• produces basic sentence forms and some correct simple sentences but subordinate structures are rare • errors are frequent and may lead to misunderstanding	• uses a limited range of pronunciation features • attempts to control features but lapses are frequent • mispronunciations are frequent and cause some difficulty for the listener
3	• speaks with long pauses • has limited ability to link simple sentences • gives only simple responses and is frequently unable to convey basic message	• uses simple vocabulary to convey personal information • has insufficient vocabulary for less familiar topics	• attempts basic sentence forms but with limited success, or relies on apparently memorised utterances • makes numerous errors except in memorised expressions	• shows some of the features of Band 2 and some, but not all, of the positive features of Band 4
2	• pauses lengthily before most words • little communication possible	• only produces isolated words or memorised utterances	• cannot produce basic sentence forms	• Speech is often unintelligible
1	• no communication possible • no rateable language			
0	• does not attend			

IELTS is jointly owned by the British Council, IDP: IELTS Australia and Cambridge English Language Assessment.

Page 1 of 1

Appendix 2



Speaking

This test takes between 11 and 14 minutes and is conducted by a trained examiner. There are three parts:

Part 1

The candidate and the examiner introduce themselves. Candidates then answer general questions about themselves, their home/family, their job/studies, their interests and a wide range of similar familiar topic areas. This part lasts between four and five minutes.

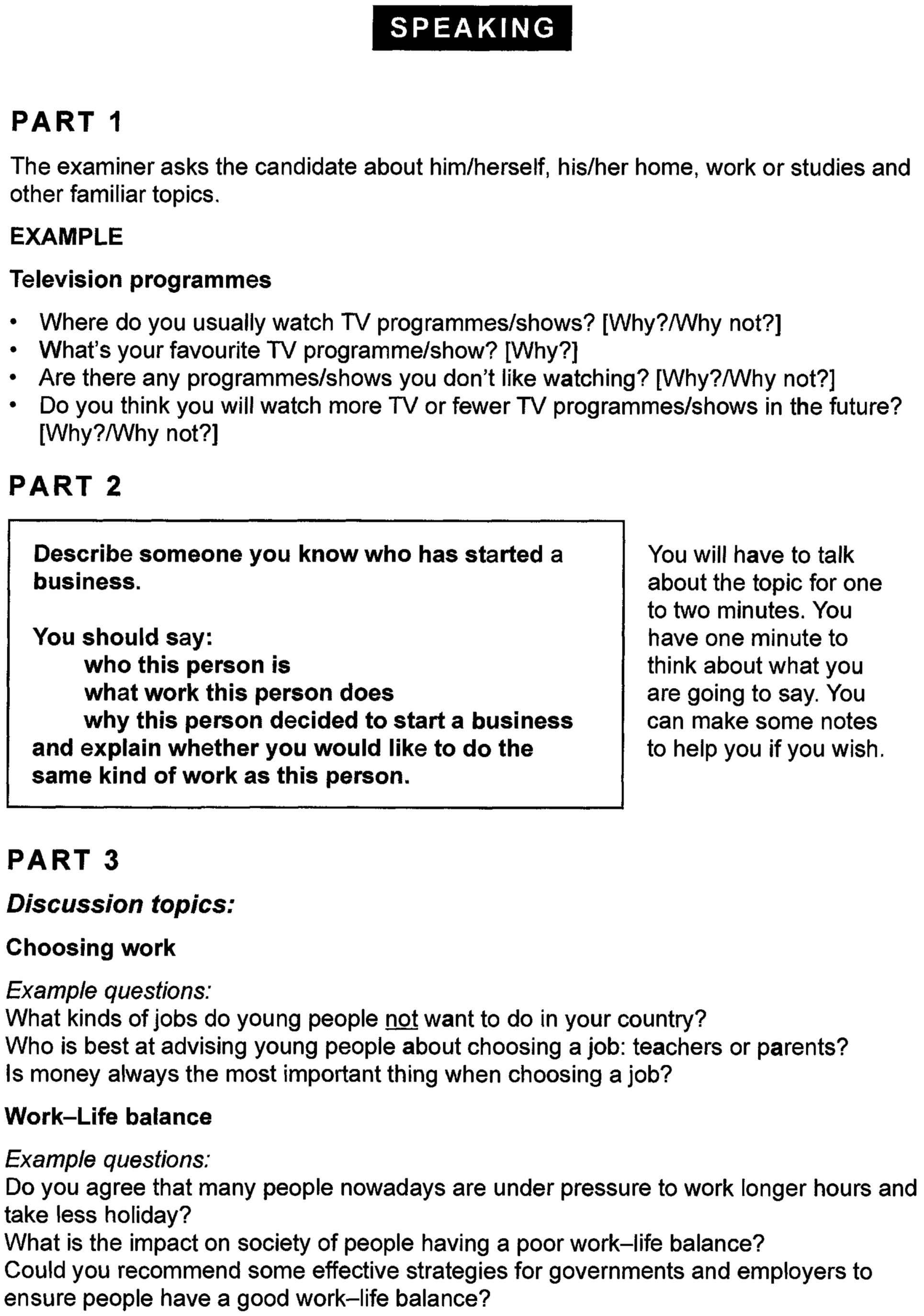
Part 2

The candidate is given a task card with prompts and is asked to talk on a particular topic. The candidate has one minute to prepare and they can make some notes if they wish, before speaking for between one and two minutes. The examiner then asks one or two questions on the same topic.

Part 3

The examiner and the candidate engage in a discussion of more abstract issues which are thematically linked to the topic in Part 2. The discussion lasts between four and five minutes.

The Speaking test assesses whether candidates can communicate effectively in English. The assessment takes into account Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy, and Pronunciation. More information on assessing the Speaking test, including Speaking assessment criteria (public version), is available on the IELTS website.



SPEAKING

PART 1

The examiner asks the candidate about him/herself, his/her home, work or studies and other familiar topics.

EXAMPLE

Television programmes

- Where do you usually watch TV programmes/shows? [Why?/Why not?]
- What's your favourite TV programme/show? [Why?]
- Are there any programmes/shows you don't like watching? [Why?/Why not?]
- Do you think you will watch more TV or fewer TV programmes/shows in the future? [Why?/Why not?]

PART 2

Describe someone you know who has started a business.

You should say:
who this person is
what work this person does
why this person decided to start a business
and explain whether you would like to do the same kind of work as this person.

You will have to talk about the topic for one to two minutes. You have one minute to think about what you are going to say. You can make some notes to help you if you wish.

PART 3

Discussion topics:

Choosing work

Example questions:
What kinds of jobs do young people not want to do in your country?
Who is best at advising young people about choosing a job: teachers or parents?
Is money always the most important thing when choosing a job?

Work–Life balance

Example questions:
Do you agree that many people nowadays are under pressure to work longer hours and take less holiday?
What is the impact on society of people having a poor work–life balance?
Could you recommend some effective strategies for governments and employers to ensure people have a good work–life balance?